AD_____

AWARD NUMBER: W81XWH-07-1-0492

TITLE: Interdisciplinary Studies on the Combat Readiness and Health Issues Faced by Military Personnel

PRINCIPAL INVESTIGATOR: John Hart, Jr., M.D.

CONTRACTING ORGANIZATION: University of Texas at Dallas
Richardson, TX 75080

REPORT DATE: September 2009

TYPE OF REPORT: Final Addendum

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1 September 2009	Final Addendum	1 Sep 2008 – 31 Aug 2009

4. TITLE AND SUBTITLE

Interdisciplinary Studies on the Combat Readiness and Health Issues Faced by Military Personnel

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0492

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John Hart, Jr., M.D.

E-Mail: jhart@utdallas.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas at Dallas
Richardson, TX 75080

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this research project was twofold: Task # 1a: Assemble multimodal human performance laboratory including complex human motor assessment system, 128 channel EEG/ERP, pupilometer/eyetracking system, and repetitive transcranial magnetic stimulation system. Task # 1b: Conduct a pilot research study demonstrating the capabilities of performing multimodal assessment of object retrieval, particularly when those objects may be considered threatening or nonthreatening. Task #1a. has been accomplished and reported previously. The pilot study (Task #1B.) has not been completed.

15. SUBJECT TERMS
rTMS, threat, PTSD, EEG, fMRI

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	19	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

INTRODUCTION:

The goal of this research project was twofold:

Task # 1a: Assemble multimodal human performance laboratory including complex human motor assessment system, 128 channel EEG/ERP, pupilometer/eyetracking system, and repetitive transcranial magnetic stimulation system.

Task # 1b: Conduct a pilot research study demonstrating the capabilities of performing multimodal assessment of object retrieval, particularly when those objects may be considered threatening or nonthreatening.

BODY:

Task #1a. has been accomplished and reported previously. The Multi-Modal Brain-Motor Performance Laboratory is fully constructed and operational at the University of Texas at Dallas Center for BrainHealth located at 2200 W. Mockingbird Lane, Dallas, Texas 75235 in room 3.120. Additional planned components (a high-end immersive driving/task simulator and an eight camera human motion capture and analysis system) were also constructed and are located at the University of Texas at Arlington's Human Performance Institute, Nedderman Hall, Room 241.

We have recently also added to the Multi-Modal Brain-Motor Performance Laboratory two motion detection cameras that with sensors attached to a subject's knees and ankles can record human gait patterns for analysis.

The pilot study (Task #1B.) has not been completed. This has been secondary to the delay in IRB approval.

KEY RESEARCH ACCOMPLISHMENTS:

Completion of the integrated Multimodal Human Performance Laboratory, including:

- Repetitive Transcranial Magnetic Stimulation System with Magstim Super Rapid package and BrainSight Software System
- SensoMotoric Instruments (SMI) Eyetracker system
- Biologics EEG System
- Human Motor Performance assessment system

We have upgraded the system further by purchasing on our own an EMG unit to integrate with the other equipment.

Personnel have preparing for the integrated studies to be performed in the multi-modal laboratory be performing the following:

1. Working on technological development in time-frequency EEG analysis which has lead to a recent publication (Ferree et al., Neuroimage, 2009).
2. Working with engineering on inspection of the rTMS unit and upgrade and integration with the EMG unit added to the multi-modal laboratory.
3. Pursued further development and evaluation of key conceptual frameworks (General Systems Performance Theory and the Elemental Model for Human Performance) and analytic techniques (e.g., Nonlinear Causal Resource Analysis) as described in our original proposal.

In addition, we have pursued development of task scenarios for use with the high –fidelity driving simulator that was acquired under this project. Specifically, we are using the capabilities of this programmable platform to develop brief "driving tasks" that are intended to stress multiple sensory, motor, and information processing subsystems. These are designed in a manner so as to be relevant to modern-day military contexts involving urban warfare. Prototype scenarios have been programmed and have undergone operational evaluations.

REPORTABLE OUTCOMES:

Task #1a. has been accomplished as previously reported.

The IRB protocol for the pilot study is pending approval at the University of Texas Southwestern Medical Center's (UTSW) IRB for approval prior to submission to USAMRMC HRPO. We are optimistic that this IRB submission will meet with UTSW, UTD, and USAMRMC HRPO approval and be successful so our pilot study can begin. Upon its approval by UTSW and UTD IRB and USAMRMC HRPO this pilot study will be performed and completed.

Submission of a letter of intent by Dr. John Hart to CDMRP DMRDP entitled "Novel Treatment of Emotional Dysfunction in PTSD".

Submission of a letter of intent by Dr. John Hart to CDMRP DMRDP entitled "Dopamine Agonist Therapy for Semantic Memory Deficits in TBI".

Submission of a letter of intent by Dr. John Hart to USAMRMC Broad Agency Announcement entitled "Mobile Cognitive and Sleep Assessment Monitors for Battlefield Deployment".

Submission of white paper by Dr. George Kondraske to the National Security Science and Engineering Faculty Fellowship Program entitled "Performance Theoretic and Elemental Resource Modeling Approach to DOD Human Performance and Systems Integration Challenges".

Early in 2009, Dr. Kondraske was invited to participate in a DoD Task Force on Human-Systems Integration, charted by the Defense Safety Oversight Council (DSOC) and chaired by USAF Major General Thomas Travis. One outcome of a meeting of this task force was to establish as a major objective "fertilization of research and development of GSPT (General Systems Performance Theory) throughout relevant DoD entities."

Kondraske, G.V. (2009) "Big Picture Frameworks: General Systems Performance Theory and the Elemental Resource Model for human performance", DSOC Task Force on Human-Systems Integration, Kick-off Meeting, April 2, Monterrey, CA.

Kondraske, G.V. (2009) "General Systems Performance Theory and the Elemental Resource Model for human performance, Naval Postgraduate School, April 2, Monterrey, CA.

Kondraske,G.V. and Stewart, R.M. (2009) New methodology for identifying hierarchical relationships among performance measures: Concepts and demonstration in Parkinson's Disease. CD-ROM Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, Sept 2 – 6, 5279-5282.

Armstrong, J.D., Kondraske, G.V., and Stewart, R.M. (2009, in press) A composite steadiness (tremor) measure reflecting full (6 DOF) motion. 2009 Annual Conference of the National Center for Human Performance, Houston (with poster presentation, Nov 6, 2009).

Kondraske, G.V., Mathew, S.J. and Stewart, R.M. (2009, in press) Measurement of Visual-auditory information processing dexterity: A basic situational awareness performance resource. 2009 Annual Conference of the National Center for Human Performance, Houston (with poster presentation, Nov 6, 2009).

Kondraske, G.V. (2008) Establishing the relationship between basic subsystem performance capacities and performance in functional tasks. 2008 Annual Conference of the National Center for Human Performance, Houston (with poster presentation, Nov 7, 2008).

CONCLUSION:

In summary, the assimilation of this unique multimodal human cognitive-motor assessment laboratory is a significant accomplishment and will allow for numerous innovative studies. We plan to perform the proposed

study after IRB approval and have multiple other studies for the future, including already targeted grant proposals.

REFERENCES:

Ferree, T., Brier, M., Hart, J., Kraut, M. Space-time frequency analysis of EEG data using within-subject statistical tests followed by sequential PCA. Neuroimage, 45(1):109-21, 2009.

NeuroImage 45 (2009) 109–121

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Space–time–frequency analysis of EEG data using within-subject statistical tests followed by sequential PCA

Thomas C. Ferree [a,*], Matthew R. Brier [b], John Hart Jr [b,c], Michael A. Kraut [d]

[a] Department of Radiology, University of Texas Southwestern Medical Center, Dallas, Texas, TX 75390-8896, USA
[b] Center for Brain Health, School of Behavioral and Brain Sciences, University of Texas at Dallas, USA
[c] Department of Neurology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[d] Department of Radiology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

ARTICLE INFO

Article history:
Received 7 May 2008
Revised 7 August 2008
Accepted 5 September 2008
Available online 30 September 2008

ABSTRACT

A new method is developed for analyzing the time-varying spectral content of EEG data collected in cognitive tasks. The goal is to extract and summarize the most salient features of numerical results, which span space, time, frequency, task conditions, and multiple subjects. Direct generalization of an established approach for analyzing event-related potentials, which uses sequential PCA followed by ANOVA to test for differences between conditions across subjects, gave unacceptable results. The new method, termed STAT-PCA, advocates statistical testing for differences between conditions within single subjects, followed by sequential PCA across subjects. In contrast to PCA-ANOVA, it is demonstrated that STAT-PCA gives results which: 1) isolate task-related spectral changes, 2) are insensitive to the precise definition of baseline power, 3) are stable under deletion of a random subject, and 4) are interpretable in terms of the group-averaged power. Furthermore, STAT-PCA permits the detection of activity that is not only different between conditions, but also common to both conditions, providing a complete yet parsimonious view of the data. It is concluded that STAT-PCA is well suited for analyzing the time-varying spectral content of EEG during cognitive tasks.

Introduction

Cognitive experiments involve stimuli delivered to the subject, and responses generated by the subject, in time frames that depend upon the task. Brain responses that involve increased synchrony with a consistent phase relationship to an event, e.g., stimulus or response, contribute to the event-related potential (ERP). Brain responses that involve decreased synchrony, or increased synchrony without a consistent phase relationship to an event, are best detected with time–frequency analysis. In cognitive tasks with long duration, time–frequency analysis is expected to be more fruitful than ERP analysis. While time–frequency analysis is relatively straightforward, a considerable challenge remains to reduce and summarize the numerical results, which span space, time, frequency, task conditions, and subjects.

Principal component analysis (PCA) has been applied extensively to ERP analysis, in order to reduce the waveforms in spatial and temporal dimensions (Spencer et al., 1999, 2001). The ERP is computed in each condition, by averaging the post-stimulus time series across trials. The number of trials in each condition and the variance of the mean are not retained. Then PCA is applied sequentially to spatial and temporal dimensions, and the resulting factor scores are submitted to ANOVA to test for differences between conditions. The group of subjects, rather than repeated trials, is used as the statistical ensemble when testing for differences between conditions. This approach, which we call PCA-ANOVA, is reported to work well for ERP analysis, and is now used widely. Its main limitations are that PCA-ANOVA can test only for differences between conditions, and requires multiple subjects on which to base these tests.

When extending to time–frequency analysis, the needs for data reduction are even greater, and the very nature of the data is different. Only a few studies have used PCA together with time–frequency analysis. Bernat et al. (2005) combined time–frequency analysis with PCA, but their emphasis was a comparison between different methods of time–frequency analysis. Tenke and Kayser (2005) studied the effects of transforming the power spectrum, and using an explicit reference versus the surface Laplacian. Our findings support this previous work, but neither group addressed the key question of how best to integrate PCA with statistical testing.

Following the approach established in the ERP literature, we applied PCA sequentially to frequency, space, and time dimensions, then submitted the resulting scores to ANOVA to test for differences between conditions. The results were found to be unstable, and changing the order of dimensions did not resolve the problems. We suspected that PCA, when applied first, was unable to isolate task-related changes, because the power spectrum is dominated by

* Corresponding author.
E-mail address: tom.ferree@gmail.com (T.C. Ferree).

1053-8119/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2008.09.020

features that may not be task related. If the task-related effects do not contribute the greatest variance to the matrix passed to PCA, then PCA will not isolate the task-related activity in the highest components. Because Varimax rotation behaves poorly with many components, and because the overall goal is data reduction, it is important to organize the analysis so that task-related activity appears in the first few principal components.

We hypothesized that a better approach would put statistical testing at the beginning of the analysis, in order to isolate task-related variance in single subjects. Statistical testing for differences between power spectra is standard in a large body of work on event-related synchronization (ERS) and de-synchronization (ERD), in which these tests are conducted in single subjects (Pfurtscheller and Lopes da Silva, 1999; Delorme and Makeig, 2004). The number of trials in each condition, and the variance in the estimate of the mean, are used to test statistical significance. For visualization purposes, differences that are not statistically significant are often rounded to zero. In the new approach called STAT-PCA, we tested for differences between conditions in single subjects, then followed with PCA for data reduction, and found that results were highly stable.

Within-subject statistical testing also solves several other problems inherent in the analysis of cognitive data with PCA. First, it permits testing for differences not only between conditions, but also between a given condition and baseline. We use this fact to reveal activity that is common to both conditions. Second, it has been noted that the rotation ambiguity of PCA factors may result in misallocation of variance (Wood and McCarthy, 1984), giving linear combinations of activity in the two conditions (Dien, 1998). By conducting statistical tests within subjects, activity that is different between conditions, and activity that is common to both conditions, are separated before decomposing with PCA, so an important caveat of PCA is eliminated. Third, statistical testing in single subjects isolates task-related activity in single subjects, and this facilitates clinical diagnosis in which single subjects are the focus of investigation.

Methods

Participants

The subjects were 25 young adults between the ages of 18 and 29. All were right-handed, and 10 were male. All were free from neurological or psychiatric disorders by self-report. Written informed consent was obtained from each subject prior to testing. This study was conducted according to the Good Clinical Practice Guidelines, the Declaration of Helsinki, and the U.S. Code of Federal Regulations. Written and informed consent was obtained from all participants according to the rules of the Institutional Review Board of The University of Texas at Dallas.

Stimuli and task

The data set upon which we developed these methods is part of a continuing study of semantic memory retrieval (Slotnick et al., 2002; Assaf et al., 2006). The stimuli consisted of pairs of written words, which consisted of features of familiar objects (e.g., 'desert' and 'humps'). The subjects were to determine whether the two features combined to result in retrieving the memory of a specific object (e.g., 'camel'). Other word pairs were chosen to lead to no retrieval (e.g., 'mane' and 'wings'). Subjects were instructed that the target needed to be a specific object, not merely an association between the two words. Fifty trials comprised stimulus pairs that have been shown in previous work (Assaf et al., 2006; Brier et al., 2008) to elicit retrieval of a specific object, and 50 were non-retrieval trials. The same feature words used in the object retrieval pairs were used in the non-retrieval pairs, but were re-paired with a semantically unrelated word. Each word pair appeared on the screen for 3 s, separated by a blank screen that appeared for 2–3 s (randomized). While it is true that the researchers defined which word pairs should or should not elicit retrieval, the small number of 'incorrect' responses indicates that this task does access semantic memory. The few trials with incorrect responses were also discarded.

Data acquisition

EEG data were acquired with a 64-channel, Synamps II system (Compumedics, Inc.). Data were sampled at 1 kHz and hardware filtered at 200 Hz. Electrode impedances were typically below 5 kΩ, although some were slightly higher. An experienced EEG technician preprocessed the data manually. First, data recorded from poorly functioning electrodes were identified visually and removed. Second, eye blink artifacts were removed by a spatial filtering algorithm in the Neuroscan Edit software (Compumedics, Inc.), using the option to preserve the background EEG. This option uses the singular value decomposition of a 'clean' data segment to optimize the removal of eye blinks from the continuous data (M. Pflieger, personal communication). Third, time segments containing significant muscle artifacts or eye movements were rejected.

During acquisition, time-locking events were placed in the EEG record corresponding to the white computer screen, the onset of word-pair stimuli, and button-press responses of two types. For spectral analysis, a baseline interval was defined as 1 s prior to the stimulus. In order to study stimulus-related activity, a peri-stimulus interval was defined from −1 to 3 s. The data were epoched accordingly and exported to Matlab (Mathworks, Inc.) for further analysis. We have begun to explore the benefits of epoching relative to responses, but for brevity only peri-stimulus results are reported here.

Reference correction

The data were recorded with a reference electrode located near the vertex, which results in small amplitudes over the top of the head. In order to correct for this effect, the data were re-referenced to the average voltage at each time point, which approximates the voltage relative to infinity (Nunez, 1981). In order to minimize a known bias in the electrode-based average reference (Junghofer et al., 1999), a spline-based estimate of the average scalp potential (Ferree, 2006) was computed using spherical splines (Perrin et al., 1989). Placing the electrode cap on a realistic phantom head, the electrode coordinates were digitized (Polhemus, Inc.), and these coordinates were used to fit the splines for each subject. The integrity of the spline interpolation was confirmed visually, by comparing waveforms of arbitrarily deleted channels with the original waveforms in those channels. The integrity of the spline-based average reference was confirmed by comparing topographic maps of baseline alpha power with similar maps using the cap reference, and the electrode-based average reference. In subjects with a small number of bad electrodes, the splines were used to interpolate those electrodes, to yield a total of 62 data channels in every subject. Ensuring the same number of electrodes in all subjects facilitates the matrix manipulations in sequential PCA.

Time–frequency analysis

Throughout the peri-stimulus interval, time-dependent Fourier power spectra were estimated in 0.5-second wide windows, moving in 0.05-second steps. The time of each window was defined as the center of the nonzero data in that window. The earliest time was −0.75 s, and the latest time was 2.75 s, because the centers of 0.5-sec windows cannot reach the ends of the epoch. Fourier power spectra were computed using the *pwelch* function implemented in Matlab (Mathworks, Inc.). In each window, the time series was linearly

detrended to reduce spectral leakage from the zero-frequency bin, cosine tapered to reduce spectral leakage generally, and zero-padded to 1-s duration to achieve 1-Hz frequency resolution. Each time series was then Fourier transformed, magnitude squared, and suitably normalized to obtain the power spectral density (PSD) in units μV^2/Hz. For each condition, the result was averaged across all trials, to obtain the best statistical estimate of the PSD.

The time-averaged PSD in the baseline interval was computed two ways for comparison. In both ways, the PSD was averaged across all trials in both conditions. In the first way, the time averaging was accomplished using the Welch method, in which the 1-s baseline interval was divided into three 0.5-s windows with 50% overlap (Welch, 1967). In the second way, the time averaging was accomplished by averaging the time-varying PSD over all time points prior to the stimulus. This was done because we observed in several subjects that the power values in the Welch windows were not always representative, occasionally missing large fluctuations and leading to inaccurate estimates of the temporal mean. The relationships between the estimates of baseline power, and the effects on PCA-ANOVA and STAT-PCA, are discussed in the Results.

If the results of our analysis are to be used to explain cognitive processes in terms of neural oscillations, we should state clearly what oscillations are included in our analysis. In conventional terminology, 'evoked' activity has a precise time and phase relationship with a temporal event, while 'induced' activity does not. The ERP isolates evoked activity, but spectral analysis is needed to detect induced activity. Because evoked activity also contributes to spectral power, a direct application of spectral analysis must be considered to reflect both evoked and induced activity. Some researchers have attempted to isolate induced activity, by subtracting the ERP prior to time–frequency analysis (Kalcher and Pfurtscheller, 1995; Ding et al., 2000; Truccolo et al., 2002). In the present work, it is not our goal to distinguish evoked and induced activity, and we have not implemented any method to subtract the ERP. Our results must therefore be interpreted to include both evoked and induced activity. Because it is difficult to maintain phase locking to the stimulus for long times, however, we expect that later times are dominated by induced activity, especially at higher frequencies.

Difference-mode and common-mode responses

The standard PCA-ANOVA approach, applied to ERP waveforms, is usually arranged to test for differences between conditions (Spencer et al., 1999, 2001; Dien et al., 2003). The calculation of the ERP waveform in each condition begins with subtracting the baseline, defined as the average of the pre-stimulus time series across time and trials. Baseline subtraction is required in ERP analyses, because EEG time series are prone to slow drift from electrode polarization and imperfect amplifier properties. Yet activity in the baseline interval may include residual activity from the previous response and/or anticipation of the next stimulus. For this reason and others, many ERP researchers prefer to focus only on differences between conditions.

In the ERD/ERS literature, it is common to quantify brain oscillations relative to baseline (Pfurtscheller and Lopes da Silva, 1999; Delorme and Makeig, 2004). In the power spectrum, any electrode or amplifier drift appears in the zero-frequency term, so this effect is not central. (An exception is that *any* power in the lowest frequency bin, e.g., less than 0.5 Hz when using 1-Hz frequency bins, may contaminate higher frequencies by spectral leakage, but this effect is minimized by detrending the data in each window before Fourier transforming.) Of course, the baseline interval may still be contaminated, by the previous response or anticipation of the next stimulus. Because the Fourier transform is squared in each trial to compute power, and because the moving windows have non-zero width determined by the taper, the same variation in the inter-stimulus interval may not be as effective.

Despite the issues inherent to defining and interpreting a baseline interval, we argue that it is compelling to consider *both* differences between conditions *and* differences relative to baseline in parallel. In the case of our semantic retrieval task, for example, it seems likely that studying the words visually, searching for associations, and launching a motor response are common to both conditions, while retrieving a word from semantic memory occurs only in one condition. Although the mental processes of semantic retrieval are not yet understood well enough to list and categorize each one, considering both kinds of activity obviously gives a more complete picture of the data. Consider a hypothetical example, in which there is greater power in condition A than condition B, for some frequency band and time interval. From this information alone, any of the following could be true: 1) condition A has ERS that condition B does not, 2) condition B has ERD that condition A does not, 3) conditions A and B both have ERS, but condition A has more, 4) conditions A and B both have ERD, but condition B has more. Only by studying the differences relative to baseline is it possible to distinguish these cases.

Following this logic, we define the 'difference-mode' response as the difference of moving-window power spectra between two conditions, without reference to baseline, and define the 'common-mode' response as the difference of moving-window power spectra between *both* conditions and baseline. By *both* conditions, we mean the average of the PSD in both conditions, weighted by the number of trials in each condition; that is equivalent to taking the union of the trials in both conditions before averaging across trials. The rational for pooling the responses from both conditions to form the common-mode response, rather than looking at each response separately relative to baseline, is that half of the information in the separate responses is already contained in the difference-mode response, and the common-mode response is orthogonal to that. Taken together, therefore, the difference-mode and common-mode responses give the most complete yet parsimonious view of task-related activity. We submit the results from both modes separately to PCA. Once the prominent factors from both modes are identified, these can guide inference about the behavior in each condition.

To be precise, consider the algebraic definitions of common mode (CM) and difference mode (DM), assuming the number of trials in the two conditions is equal. Let R represent the time-varying PSD in the retrieval condition, N represent the time-varying PSD in the non-retrieval condition, and B represent the time-averaged PSD in the baseline interval. At each frequency, electrode, and time point, we have $DM = R - N$, and $CM = (R+N)/2 - B$. Solving for R and N gives $R - B = CM + DM/2$, and $N - B = CM - DM/2$. Thus the response in each condition relative to baseline may be obtained trivially from the common-mode and difference-mode results. In the present work, difference-mode STAT-PCA permits a direct comparison with PCA-ANOVA, and common-mode STAT-PCA gives new information not accessible with PCA-ANOVA.

Statistical tests for differences between power spectra

In most ERP studies, only the within-subject average response is carried forward, e.g., to PCA-ANOVA, and neither the number of trials nor the variance across trials is used. Instead, the statistical ensemble used to test for differences between conditions is comprised of many subjects in the study. In our initial attempts, we adopted this approach, submitting the trial-averaged, time-varying power in each condition to three-way PCA-ANOVA. Empirically, the results were unsatisfactory (see Results). Theoretically, a disadvantage of conducting PCA first is that statistical variability in the estimates of the power spectrum, which are inherent for stochastic signals, might overly influence the PCA. On these grounds, we hypothesized that within-subject statistical tests for differences in power spectra, either between conditions or relative to baseline, would improve the results.

Statistical testing for differences between power spectra is a reasonably well-developed topic. In its simplest form, the log of the power spectrum estimate is assumed to be Gaussian distributed. The estimate based upon a finite number of samples is biased, however, by an amount that depends upon the number of samples (Thomson and Chave, 1991; Bokil et al., 2007). Using analytic expressions in these references, we corrected for this bias and estimated the variance in the mean using the appropriate number of samples in each case. The bias correction is a simple function of the number of samples, which is subtracted from the original calculation of the PSD. In the post-stimulus interval, the number of samples contributing to the PSD in each condition was taken to be twice the number of trials in that condition, accounting for sine and cosine contributions. (In the calculation of baseline power and the common mode, the number of trials was equal to the sum of the trials in both conditions.) Furthermore, in the baseline interval, because the 50% overlap combined with cosine tapering leads to approximately independent samples, the number of samples was tripled, corresponding to three 0.5-s windows in the 1-sec baseline interval.

The PSD is computed at many electrodes, frequencies, and time points. By setting α=0.05, false positives are expected at this rate. Correcting for multiple comparisons is non-trivial when the data are correlated. In the spatial dimension, data at different electrodes are correlated due to volume conduction. In the spectral dimension, spectral leakage causes neighboring frequency bins to be correlated. In the temporal dimension, overlapping windows cause neighboring time points to be correlated. One remedy in the spectral domain is to keep only contiguous sets of frequencies that are wider than the bandwidth of the analysis (Bokil et al., 2007), and another is to use the false-discovery rate (Durka et al., 2004), but neither of these approaches as published addresses false-positives in the full three dimensions of space–time–frequency. In the present work, we have chosen for simplicity not to correct for multiple comparisons. If false positives occur randomly, they should have little impact on the PCA analysis applied to multiple subjects. Indeed, the results below support the assertion that only meaningful, task-related activity emerges from STAT-PCA.

Input to principal component analysis

This work contrasts two approaches to integrating time–frequency analysis with statistical tests. In the standard method, PCA-ANOVA, the time-dependent power values in each condition are input to sequential PCA (see below) and the resulting scores are passed to ANOVA. This is a one-way (condition) repeated measures ANOVA with two levels (retrievals and non-retrievals). In this approach, the variance across trials is not retained, and the statistical ensemble used in the ANOVA is comprised of multiple subjects. In the new method, STAT-PCA, statistical tests between conditions are conducted in each subject and electrode separately, and the statistical ensemble is comprised of multiple trials. Insignificant differences are rounded to zero, in effect, eliminating noise from the results. Only the non-zero spectral differences are passed to PCA for data reduction.

Although the statistical tests for differences between spectra in STAT-PCA assumed that the log-power spectrum is Gaussian distributed, we used PSD in units of $\mu V/Hz^2$ not dB as input to PCA. This choice does not affect the set of frequencies that show significant differences, but does affect the numerical values submitted to PCA. We adopted this method after we tried both $\mu V/Hz^2$ and dB units and the former gave much better results. The log transformation renders large power values not so different from the small power values, and this had several adverse effects on the PCA results: many more factors retained, less distinctive factor loadings, and poorer agreement with group averaged data. We therefore kept the units of $\mu V/Hz^2$ as input to PCA, consistent with the recommendations of Tenke and Kayser (2005), but set the differences to zero according to the aforementioned statistical test. Even with this choice, which tends to underemphasize the power at higher frequencies, we obtained differences in the 20–35 Hz range that were reported previously in this task (Slotnick et al., 2002).

Principal component analysis

In our description of PCA, we use standard terminology and matrix organization, with some small exceptions. The data matrix is arranged with columns indexing variables and rows indexing samples. Following standard conventions in PCA analysis, the column-means are subtracted, i.e., in each column separately the mean across rows is subtracted. Following the consensus in the ERP literature (Kayser and Tenke, 2003; Dien et al., 2005), we use the covariance matrix rather than the correlation matrix. With this convention, PCA is equivalent to singular value decomposition (SVD): the right eigenvectors are called the component or factor 'loadings', and the left eigenvectors times the singular values are called the component or factor 'scores'.

In ERP analysis, PCA has been applied sequentially to reduce the results in the spatial and temporal dimensions. An important consideration is the order in which PCA is applied to the various dimensions. An early work applied PCA to the spatial dimension (Donchin, 1966), and later works applied PCA to the temporal dimension (Curry et al., 1983; Donchin and Heffley, 1979; Mocks and Verleger, 1991). More recent works applied PCA spatially then temporally (Spencer et al., 1999, 2001; Dien et al., 2003). The choice to apply spatial PCA before temporal PCA for ERP analysis was based on the argument that 'components are defined by unique patterns of scalp distributions' (Spencer et al., 2001). It has also been suggested to apply temporal PCA first (Dien and Frishkoff, 2005), because spatial components may overlap due to volume conduction. Despite all these well-reasoned arguments, no study has yet compared the effects of order in sequential PCA.

In our initial investigations, we studied PCA-ANOVA and STAT-PCA for *all six* possible orderings of frequency, space, and time. In order to keep this report tractable, we focus on one order. We arrived at this order by following one of the earlier arguments, that the best order is determined by the inherent separability of the data (Dien, 1998). First, cognitive processes are accompanied by oscillations in characteristic frequency bands. Bands have nonzero width, but do not typically overlap. This separability suggests that spectral PCA should be conducted first. Second, time–frequency analysis involves moving windows with some non-zero width, and this blurs time resolution below that of ERP analysis. This suggests that temporal PCA should be conducted last. On these arguments, we suggest that a sensible starting point is spectral–spatial–temporal PCA. Based upon our initial investigations, which spanned all six possible orderings, our impression is that this ordering produced among the most stable and sensible results for our data set.

To perform sequential PCA in this order, the data are arranged into a matrix, with columns indexing frequencies, and rows indexing the result of concatenating electrodes, time-points, conditions (PCA-ANOVA only), and subjects. First, PCA is applied to obtain the spectral loadings, and the largest factors are retained. For each spectral factor retained, the corresponding factor score is reshaped to form a matrix with columns indexing electrodes, and rows indexing the concatenation of time points, conditions (PCA-ANOVA only) and subjects. Second, PCA is applied to obtain the spatial loadings, and the largest factors are retained. For each spatial factor retained, the corresponding score is reshaped to form a matrix with columns indexing time points, and rows indexing the concatenation of conditions (PCA-ANOVA only) and subjects. Third, PCA is applied to obtain the temporal loadings, and the largest factors are retained. In PCA-ANOVA, the temporal scores are submitted to ANOVA. In STAT-PCA, the temporal scores are simply kept as the final result.

Factor retention

The main objective of PCA is dimension reduction, but determining the precise number of components or factors to retain is notoriously difficult (Zwick and Velicer, 1986; Hayton et al., 2004). All of the standard methods are based upon the eigenvalues. The goal is to identify a small set that capture most of the data variance, and are distinct from the remaining set of smaller eigenvalues. In the simplest method, the eigenvalues are plotted in a 'scree' plot, and the cut point is determined by eye (Cattell, 1966). Despite its wide use, this method is highly subjective, especially when the eigenvalues decrease gradually, or when there are several distinct steps. In the present data, we often have one very large eigenvalue, followed by one or more visible steps. To make the choice less subjective, we implemented two statistical techniques that should bracket the best choice.

The first technique is the maximum profile likelihood (MPL), which aims to determine the cut point that leads to the most natural grouping of eigenvalues (Zhu and Ghodsi, 2006). After exploring this technique extensively, we have arrived at the following impressions. When a few eigenvalues are large and similar, standing well above the others, MPL picks these few. When the first eigenvalue is much larger than the others, MPL tends to pick it, even if there is a second elbow in the scree plot just a few points away. In this way, MPL appears either to perform well, or to underestimate the number of factors. MPL has the advantage of being very efficient computationally.

The second technique is parallel analysis (PA), which is based upon rejecting the null hypothesis that the eigenvalues are the same as those of a random matrix with the same dimensions and distribution of values (Horn, 1965). Comparison studies using model data agree that this is the most reliable technique (Zwick and Velicer, 1986; Hayton et al., 2004), although it is not widely used. One study showed that, when PA is wrong, it tends to overestimate the number of factors to retain (Zwick and Velicer, 1986), although another study points out that PA tends to underestimate the number of factors when the first eigenvalue is large (Turner, 1998). Despite having a single large eigenvalue in our data, we have found that PA always estimates more factors than MPL.

In order to generate random matrices for PA, we shuffled the values in the original matrix and computed the eigenvalues using identical procedures. For each eigenvalue generated from the null distribution, we computed the mean across the 100 random matrices. Eigenvalues from the real matrix that were greater than the mean eigenvalue from the random matrices were considered descriptive of the covariance structure of the data. The use of the mean has precedent (Hayton et al., 2004), but has also been criticized as setting the false-positive rate to 0.5. An obvious remedy is to set the false-positive rate to 0.05 (Glorfeld, 1995), but this requires many more random matrices, because evaluating the tails of a distribution is much more demanding computationally than evaluating the mean. Because PA is already quite demanding, and because a large false-positive rate corresponds to overestimating the number of factors, we used the mean of 100 random matrices as the threshold, and we interpret this PA estimate as an upper bound.

Factor rotation and refinement of factor retention

PCA decomposes a data matrix into orthogonal components. While this is often effective for separating signal and noise subspaces, it is well known that the factors retained according to relative variance alone may not provide the most useful factorization of the signal. For this reason, it is normal to apply an additional transformation, such as factor rotation, to satisfy some additional constraint. In the present work, we focus on Varimax rotation, which aims to simplify the structure of each factor by loading its variance onto the smallest number of elements. In applications to ERP data, there is general agreement that Varimax rotation helps separate distinct cognitive components (Kayser and Tenke, 2003; Dien et al., 2005).

As described in the previous subsection, the eigenvalue-based techniques MPL and PA are helpful in determining the number of factors to retain, but they leave some ambiguity and subjectivity in the choice. In the present work, we have made inroads toward an improved method for factor retention. It is based upon the common understanding that retaining too many factors becomes problematic when using rotation (Dien et al., 2007). Our idea is to compare the *rotated* factor loadings with the data, as a function of the number of factors retained. In order to compare the factor loadings with the data, we need some measure of the data that has the same dimension as the factor loadings. Because the column-means are subtracted prior to PCA, and because PCA is fundamentally a variance-based technique, the simplest non-zero data measure is the column-variance. Analogous to the definition of column-mean (see above) the column-variance is the variance of each column using the rows as samples. Of course, the column-variance of the data matrix contains less information than the full covariance matrix, but still provides a sensible data measure for evaluating the rotated factors.

To apply this idea in practice, in each step of sequential PCA, we plotted the original and rotated eigenvectors along with the normalized column-variance (see Fig. 3). Varying the number of factors retained within the range bracketed by MPL and PA, we retained the minimum number of factors necessary to explain the prominent features in the data. This procedure was carried out to decide the number of factors to retain in each step of sequential PCA. The choice to limit the number of factors to the prominent features in the column-variance, within the plausible range bracketed by MPL and PA, amounts to a conservative choice of the number of factors, because we know that the full covariance matrix contains more information. This use of the eigenvectors in conjunction with the eigenvalues to make choices about factor retention deserves more development, but even in its present form it is evident that using the rotated eigenvectors in this way is more rigorous than using the eigenvalues alone.

Factor visualization and interpretation

Decisions about factor retention define a *factor tree* that associates each spectral factor with one or more spatial factors, and each spatial factor with one or more temporal factors. Each set of spectral, spatial, and temporal factors defines a *factor triplet*. Spectral and temporal loadings are plotted as two-dimensional curves, and spatial loadings are plotted as topographic maps. The loadings are normalized, but the signs are arbitrary. In order to visualize the spectral and spatial loadings consistently, we flipped their sign if their maximum absolute value corresponded to a negative value. Most spectral loadings had a single peak, although sub-peaks are often present. Most spatial loadings were either focal or bi-focal maps, appearing to represent semi-localized oscillators. Most temporal loadings were less compact, and the choice of orientation was usually arbitrary. In PCA, the (column-mean subtracted) data matrix is approximated as a product of spectral, spatial, and temporal loadings. In our conventions, any negative signs were effectively assigned to the temporal loadings as follows.

In order to determine the most sensible sign for the temporal loadings, we compared them to the subject-averaged data. For each triplet, we selected the peak frequency of the spectral factor, and one or more peak electrodes of the spatial factor. For these peaks, we averaged the common-mode and difference-mode PSD values across subjects. The temporal factors were flipped and scaled to have maximum similarity with the subject-averaged PSD. Visualization of the temporal loading along with the group-averaged data confirms the internal validity of the process, and permits each factor to be interpreted as ERD versus ERS.

Metric for component similarity

This paper demonstrates that STAT-PCA performs better than PCA-ANOVA in several ways, and showing that requires comparing PCA factors quantitatively. Each *factor triplet* is comprised of a spectral loading F, a spatial loading S, and a temporal loading T. For two triplets: (F_a, S_a, T_a) and (F_b, S_b, T_b), we define the *triplet similarity metric*

$$\Gamma_{ab} = (F_a^T F_b)(S_a^T S_b)(T_a^T T_b)$$

where ^{T} denotes vector transpose. $\Gamma_{ab}=1$ implies a perfect match, and $\Gamma_{ab}=0$ implies perfect orthogonality. Because small variations in the data can change the ordering of factors with similar eigenvalues, it is important to consider all possible orderings. Consider the general case, in which set A has M triplets, and set B has N triplets. To compare all pairings of factors from sets A and B, we constructed the matrix Γ_{ab}, for $a=1,\ldots, N$ and $b=1,\ldots, M$. We summarized this matrix by computing the maximum of each row, to identify for each factor in set A the most similar factor in set B, independent of order. In order to generate a summary statistic across subjects, we took the maximum along the largest dimension of Γ (e.g., if $a>b$ then the maxima were taken across the rows). This vector was then averaged to obtain a single global metric for each subject. These values were subjected to a paired t-test to determine if STAT-PCA was in fact more stable than PCA-ANOVA.

Results

Time-varying power and mode transformation in single subjects

Fig. 1a shows the time-varying power for a single subject, electrode OZ, frequency 10 Hz, relative to the Welch baseline power. Both conditions show power fluctuations during the baseline interval, decreased power after stimulus presentation, and a return toward baseline starting near 1 s. Fig. 1b shows the difference-mode (dashed) and common-mode (solid) responses. The common mode captures the strong decrease that is common to both conditions. The difference mode is much smaller, although some deflections are visible.

Fig. 1a also shows with dots the time points at which the power in each condition was significantly different from baseline ($p<0.05$). Neither condition was significantly different from baseline prior to the stimulus. Fig. 1b shows with dots the time points at which the power in each mode was significantly different from baseline ($p<0.05$). The common mode (solid dots) was significantly different from baseline only after the stimulus. The difference mode (open dots) showed significant differences sporadically, including some points prior to the stimulus. Because the stimuli were randomized, there can be no systematic difference in baseline between the two conditions. We conclude that the differences between conditions prior to the stimulus are due to random variations, and the points that passed the statistical test in this interval are false positives. Because false positives occur randomly across subjects, however, they are not expected to influence the PCA applied to multiple subjects.

Results of automated tests for factor retention

Fig. 2 shows spectral eigenvalue plots computed for PCA-ANOVA and STAT-PCA in difference mode. The filled dots show MPL and PA recommendations for the number of retained factors. In all cases shown, MPL=1 or 2, and PA>MPL. Fig. 2a was found for PCA-ANOVA without baseline subtraction. The second factor represents 60 Hz noise, as described below. Fig. 2b was found for PCA-ANOVA with baseline subtraction. Even though the MPL and PA values for factor retention in Fig. 2b are identical to those in Fig. 2a, baseline subtraction moves 60 Hz noise from the second factor to beyond the fifth factor. Fig. 2c was found for difference-mode STAT-PCA, and Fig. 2d was found for common-mode STAT-PCA. Broadly speaking, the eigenvalue plots for STAT-PCA are similar to PCA-ANOVA, featuring 1–2 large values, followed by a small set preceding an elbow in the range 5–8. The values for MPL and PA bracket any visible elbow, except in this example of common-mode STAT-PCA. Further exploration of factor retention choices in common-mode STAT-PCA showed that, although there appears to be a step after five factors, only two factors were interpretable.

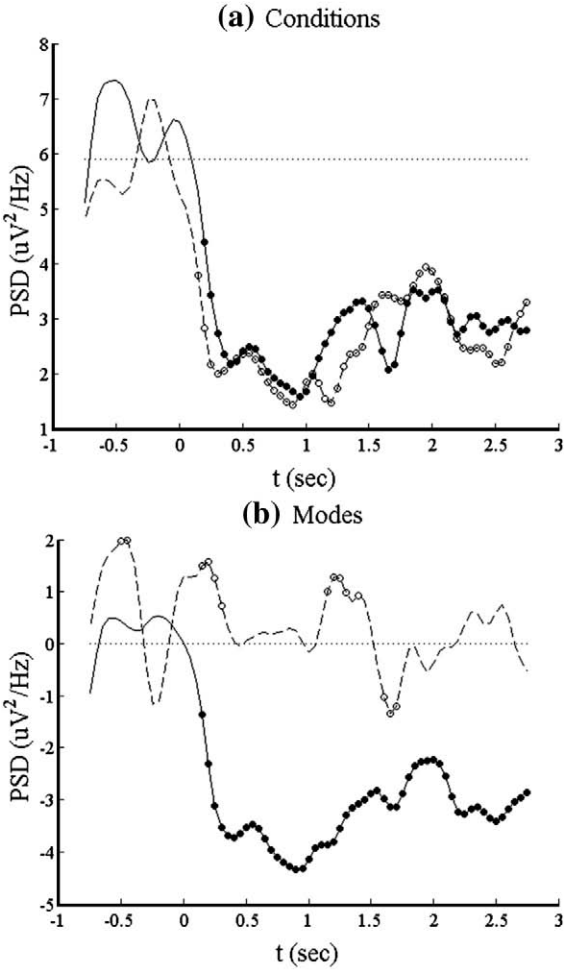

Fig. 1. Time-varying power at 10 Hz in electrode OZ for single subject: (a) retrieval condition (solid), non-retrieval condition (dashed), baseline (dotted); (b) difference mode (dashed), common mode (solid), zero (dotted). Dots show statistical significance for $p<0.05$.

Refined factor retention using factor rotation

The MPL and PA algorithms provide useful guidelines for the number of factors to retain. In order to select a number within this range, we used additional information about how the rotated eigenvectors correspond with the column-variance of the data matrix. Fig. 3 shows the un-rotated (thin dashed lines) and rotated (thin solid lines) along with the column-variance (thick solid line), for difference-mode spectral STAT-PCA, and four choices of the number of retained factors K. For $K=1$, the eigenvector (black) captures the low-frequency behavior of the column-variance, but ignores the narrow alpha peak and broad beta peak. For $K=2$, the first eigenvector (black) behaves similarly, and the second eigenvector (blue) captures the alpha peak.

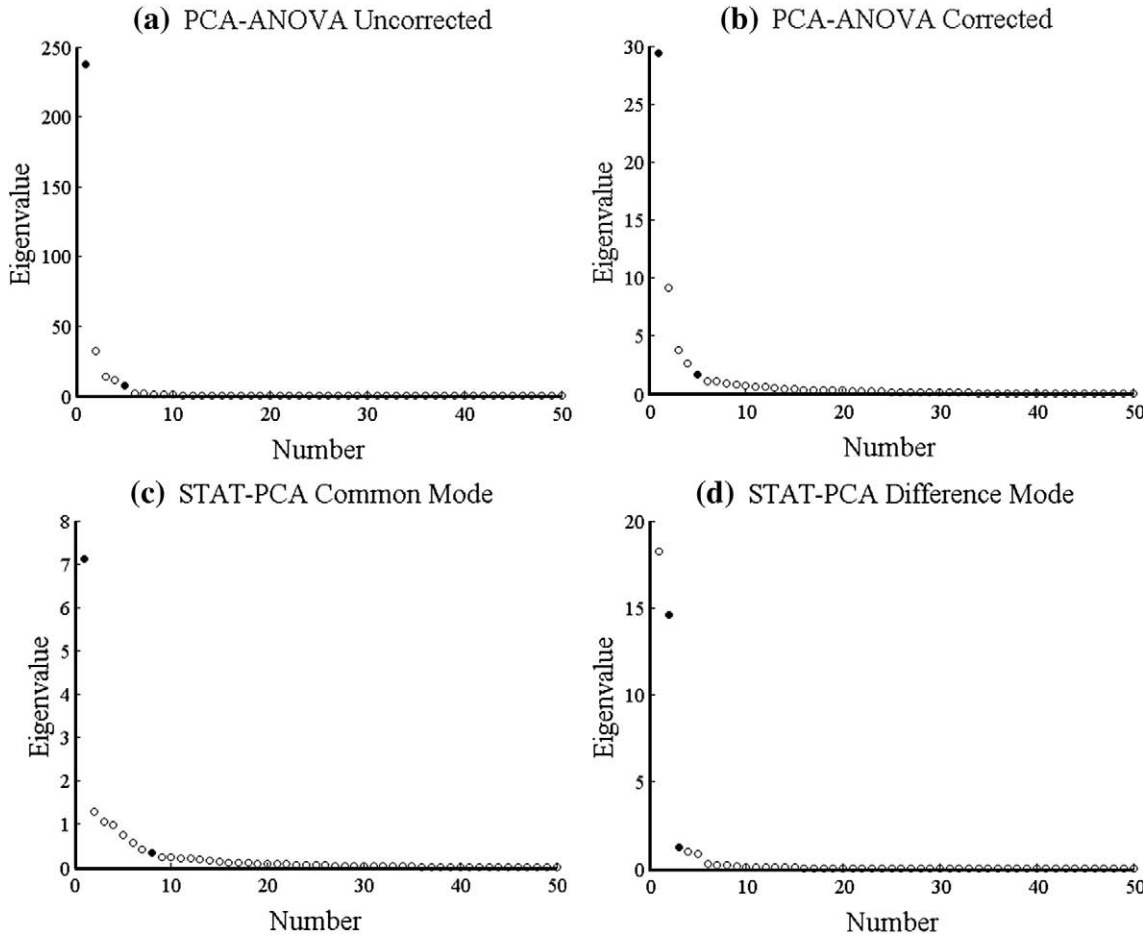

Fig. 2. Scree plots for spectral PCA: (a) PCA-ANOVA without baseline subtraction, (b) PCA-ANOVA with baseline subtraction, (c) STAT-PCA in difference mode, and (d) STAT-PCA in common mode. Filled dots indicate MPL and PA choices for factor retention; in all cases MPL<PA.

The effects of rotation are minimal in this case. For $K=3$, the roles of the first two eigenvectors are unchanged, and the third eigenvector (red) is quite complicated, including peaks at 1 Hz, 4 Hz, 11 Hz, and 29 Hz, with non-uniform signs. Again the effects of rotation are minimal. For $K=4$, the roles of the first two eigenvectors are again unchanged. Although the third eigenvector (red) is still complicated, it is simplified slightly by losing much of its peak at 29 Hz after rotation. The fourth eigenvector (green) isolates 29 Hz almost exclusively after rotation. We used these arguments to support the choice of $K=4$ in this example, and note that this choice is well within the upper limit recommended by PA (Fig. 2c). A similar procedure was used to select K for each step of sequential PCA.

Retention of 60 Hz noise in PCA-ANOVA

In our initial explorations with PCA-ANOVA we passed the time-varying PSD in both conditions to PCA, without baseline subtraction. We kept frequencies up to 100 Hz to see if any high-gamma activity could be found. We found that without baseline subtraction one of the largest components (second spectral factor, fifth spatial factor, third temporal factor) that survived the ANOVA ($F(1,48)=5.74$; $p=0.0205$) reflected 60 Hz noise. The spectral loading had a single, narrow peak at 60 Hz. The spatial loading was peaked near the ground electrode, consistent with theory (Ferree et al., 2001). The time course of the temporal factor was non-descript. We had not anticipated this result, because ANOVA was supposed to isolate differences between conditions, and the difference in 60 Hz noise between conditions should be negligible.

An explanation for why this happens is as follows. When conducting spectral PCA, the data matrix has rows that include electrodes, time points, conditions (PCA-ANOVA only) and subjects. Prior to forming the covariance matrix, the column-means are subtracted from the data matrix, consistent with the usual definition of covariance matrix. The column-variances are generally non-zero, however, even for the 60 Hz column. Because this feature dominates the covariance matrix, 60 Hz noise emerges as one of the largest factors. While it may be possible to dismiss this, saying that 60 Hz noise is easily filtered, or that one could simply limit the frequency range to 1–50 Hz, we view this as an indication that PCA-ANOVA failed to isolate task-related differences between conditions. Attempting to remedy this problem, we subtracted the baseline power from the moving-window PSD in each condition, in analogy with ERP analysis. We found that the 60 Hz factor no longer reached significance in the ANOVA, and tentatively concluded that we had solved this problem.

Sensitivity of PCA-ANOVA to the definition of baseline power

We considered two ways of estimating the baseline power spectrum. The first way was based upon the Welch method, using 50% overlapping windows. In the 1-sec baseline interval, this gives three 0.5-s windows that are nearly independent with the cosine tapering. As shown in Fig. 1, power varies in the baseline interval. The second way was based upon averaging the time-dependent power in all moving windows (0.05 s steps) within the baseline interval. The two ways of computing the baseline power give slightly different results, because the Welch estimate can be seen as three samples among many obtained with sliding windows. Fig. 4 shows scatter plots comparing the two ways of computing the baseline power in a single subject (a) and all subjects (b). In the single-subject case, for this

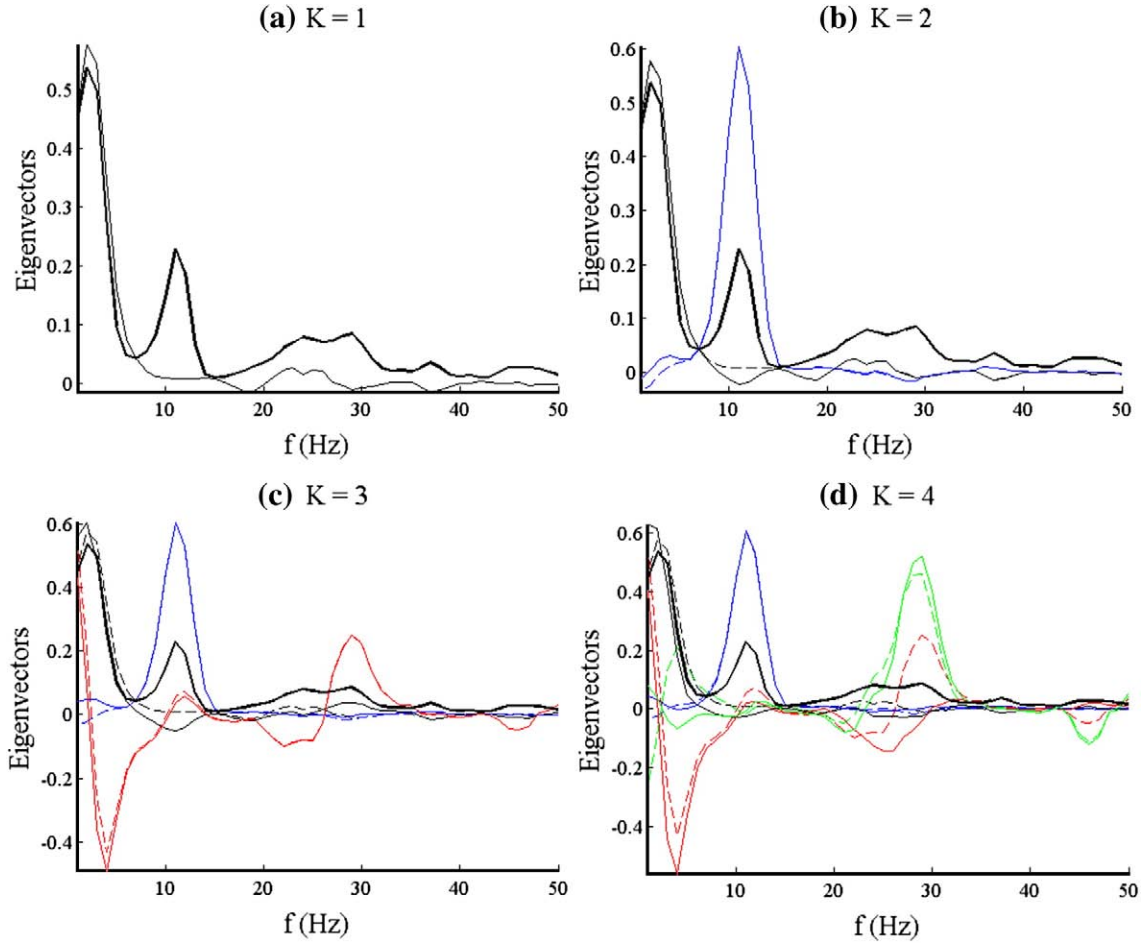

Fig. 3. Illustration of strategy for factor retention, shown for spectral STAT-PCA in difference mode. The thick solid curve represents the column-variance of the data array. Thin dashed curves represent eigenvectors before rotation. Thin solid curves represent eigenvectors after rotation. As more factors are retained, their eigenvectors are colored as follows: first (black), second (blue), third (red), fourth (green).

particular subject, the two estimates are highly correlated, suggesting small variability in power during the baseline interval. In the all-subject case, some values deviate far from linearity, revealing large baseline variability in some subjects. Further analysis revealed that these deviants come from one subject mainly, and 2–3 others secondarily. Yet these subjects are not 'bad' per se, as their EEG appear fine, and their responses to the stimulus are visible.

In PCA-ANOVA, the two methods of computing baseline power, Welch 50%-overlap estimate and sliding-window 90%-overlap estimate, gave three factors. Table 1 shows the triplet similarity matrix

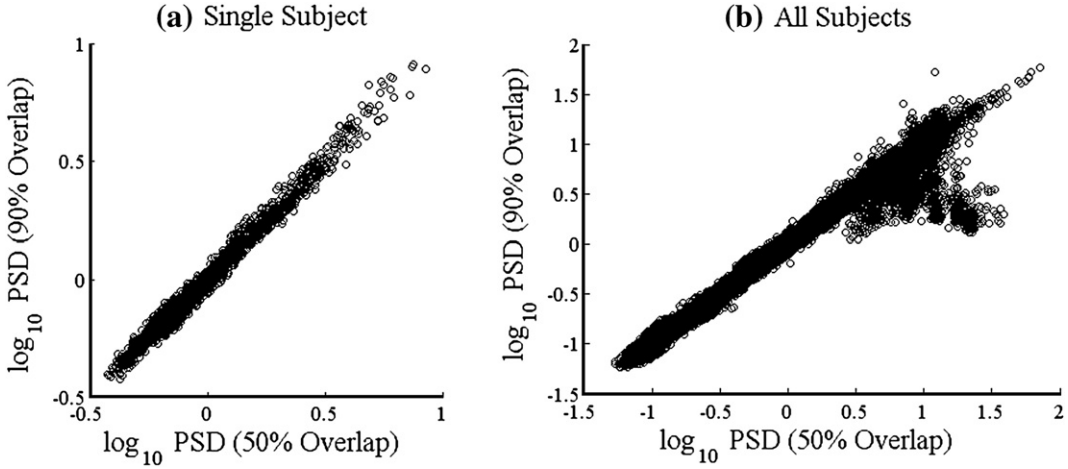

Fig. 4. Comparison of baseline power in two methods of computing for (a) single subject, (b) all 25 subjects. Horizontal axis shows results for 50% overlap (Welch method). Vertical axis shows results for 90% overlap (averaging all values in moving-window PSD estimate).

Table 1
PCA-ANOVA: baseline power

		50% Overlap		
		Triplet 1	**Triplet 2**	**Triplet 3**
90% Overlap	**Triplet 1**	0.0004	0.0000	**0.0713**
	Triplet 2	0.0022	**0.0433**	0.0137
	Triplet 3	**0.0216**	0.0234	0.0081

Triplet similarity metrics for two estimates of baseline power in PCA-ANOVA. Both estimates (50% overlap, 90% overlap) gave three significant triplets.

that compares the results. We found $\max(\Gamma_{ab})=0.0713$, which implies poor correspondence between the two solutions. We conclude that PCA-ANOVA is highly sensitive to the definition of baseline power.

In STAT-PCA, taking the difference between conditions and keeping only significant differences in single subjects had the effect of removing 60 Hz activity entirely. Even when concatenating across subjects, therefore, STAT-PCA is not affected by 60 Hz. The table for STAT-PCA that would be analogous to Table 1 has ones on the diagonal and zeros elsewhere. We conclude that STAT-PCA is relatively insensitive to the definition of baseline power.

Because the definition of baseline interval, and the method of averaging used to compute baseline power, are rather arbitrary decisions made by the researcher, with little information available to confirm the absolute validity of one choice over another, we argue that any method for analyzing group data should be minimally sensitive to these kinds of choices. The lack of robustness of PCA-ANOVA to the definition of baseline power raises serious concerns about reproducibility of results obtained with this method. In contrast, STAT-PCA isolates task-related spectral changes reliably, which is the stated goal of this entire analysis.

Sensitivity of PCA-ANOVA to the deletion of a single subject

When looking for a group effect, it is generally undesirable for one subject to influence the results excessively. In order to assess stability of the factors retained, we calculated PCA-ANOVA for the entire group and compared the results with those obtained by deleting each subject one at a time. When $N=25$ subjects were used, four factors were deemed significant by ANOVA. When a single subject was removed, three factors were found. That alone raised concern. Furthermore, of the three triplets found when $N=25$, only one of those triplets was found when $N=24$. We emphasize that the subject removed was *not* one of the subjects that exhibited high baseline variability in Fig. 4b. Indeed, Table 2 was recomputed for each of the 25 subjects separately, and similar results were obtained. To quantify this, we generated the summary statistics as described above and the mean was found to be 0.6870. This implies that PCA-ANOVA is very sensitive to the deletion of a random subject, making it difficult to generate reproducible results.

Robustness of STAT-PCA to the deletion of a single subject

In exact parallel to the procedure used to remove a single subject in PCA-ANOVA, single subjects were removed one at a time from the STAT-PCA analysis and the results compared to the group. In order to make the most direct comparison with PCA-ANOVA, only difference-mode STAT-PCA is shown. Table 3 shows that six triplets were retained for both $N=25$ and $N=24$. Most importantly, the matrix is very nearly the identity matrix, i.e., not only were the same triplets obtained, but they were retained in the same order. Table 3 was recomputed for each of the 25 subjects separately, and similar results were obtained for nearly all subjects. To quantify this, we generated the summary statistics as described above, exactly as they were calculated for PCA-ANOVA. The mean value was 0.9648. The values obtained from PCA-ANOVA and STAT-PCA were compared using a one tailed t-test and found to be significantly different ($t(24)=1.8655$; $p=0.0371$). This implies that STAT-PCA is highly stable to deletion of a random subject, which helps insure reproducible results.

Task-related factors in STAT-PCA

For our group of 25 subjects, STAT-PCA produced two common-mode (CM) triplets and six difference-mode (DM) triplets. Because the goal of this paper is to illustrate the methodology, and establish the internal consistency of the principle components with respect to the original data, only three examples are presented here. A thorough description of all the components and their interpretation as cognitive processes will be presented elsewhere. For each triplet below, thick curves represent factor loadings, and thin curves represent the group-averaged PSD at particular frequencies and electrodes as noted. In topographic plots, the color scale is dark for small values, red for intermediate values, and white for large values. The display of electrodes on the head model is slightly compressed, so that inferior occipital electrodes are also visible.

Fig. 5 shows common-mode triplet CM (1,1,1). The spectral loading has a prominent peak at 11 Hz, with a smaller peak near 2 Hz. The spatial loading is spread over occipital–parietal cortex, and is slightly left-lateralized to peak at electrode PO3. The temporal loading is shown with the group-averaged, common-mode PSD at 11 Hz and electrode PO3 (thin line). Both the loading and data start near zero, fall abruptly until 1 s, then return toward baseline through the rest of the epoch. The temporal loading matches the group-averaged data remarkably well, confirming the internal validity of this factor, and showing that this triplet reflects ERD not ERS.

Fig. 6 shows difference-mode triplet DM (4,1,1). The spectral factor has a single, prominent peak at 29 Hz. The spatial topography has two distinct peaks, bilaterally distributed in frontal electrodes F5 (left) and

Table 2
PCA-ANOVA: subject deletion

		$N=25$		
		Triplet 1	**Triplet 2**	**Triplet 3**
$N=24$	**Triplet 1**	0.0000	0.0000	0.0006
	Triplet 2	**0.9993**	0.0004	0.0001
	Triplet 3	0.0005	**0.0934**	0.0000
	Triplet 4	0.0001	0.0005	**0.0045**

Triplet similarity metrics for entire subject group ($N=25$) and one subject deleted ($N=24$) in PCA-ANOVA. Calculations for $N=24$ gave a different number of significant triplets depending upon the subject that was dropped.

Table 3
STAT-PCA: subject deletion

		$N=25$					
		Triplet 1	**Triplet 2**	**Triplet 3**	**Triplet 4**	**Triplet 5**	**Triplet 6**
$N=24$	**Triplet 1**	**0.9999**	0.0052	0.0000	0.0022	0.0074	0.0014
	Triplet 2	0.0053	**0.9999**	0.0002	0.0081	0.0022	0.0000
	Triplet 3	0.0000	0.0002	**1.0000**	0.0002	0.0003	0.0001
	Triplet 4	0.0023	0.0079	0.0002	**0.9996**	0.0170	0.0004
	Triplet 5	0.0071	0.0023	0.0003	0.0168	**0.9993**	0.0016
	Triplet 6	0.0014	0.0000	0.0001	0.0004	0.0016	**1.0000**

Triplet similarity metrics for entire subject group ($N=25$) and one subject deleted ($N=24$) in STAT-PCA. Calculations for $N=24$ gave six significant triplets consistently.

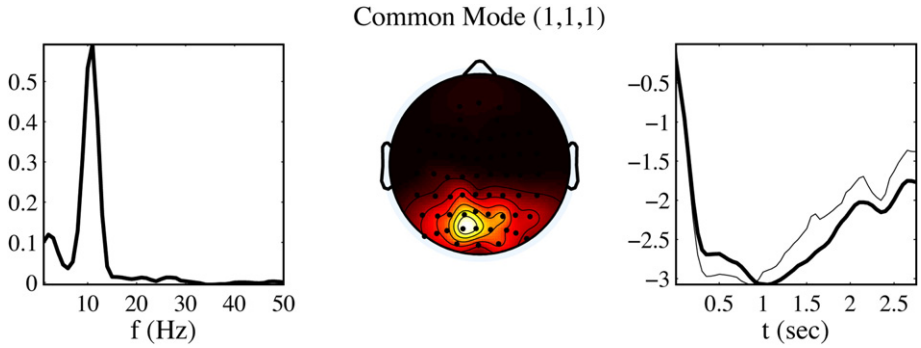

Fig. 5. STAT-PCA triplet CM (1,1,1). The temporal factor (thick line) is plotted with the group-averaged PSD at 11 Hz and electrode PO3 (thin line).

AF4 (right). Because the display of the electrodes is slightly compressed, F5 is actually more lateral than indicated in the graph. The temporal factor (thick solid line) is shown with the group-averaged, difference-mode PSD at 29 Hz, for electrodes F5 (thin solid line) and AF4 (thin dashed line). The time course of this factor and both electrodes are nearly identical. This supports the validity of finding this bifocal map, and suggests functional coupling between these locations.

Fig. 7 shows difference-mode triplets DM (1,1,1) and DM (1,1,2). The spectral factor is peaked at 1 Hz, falling to zero by 9 Hz. This low-frequency power is not merely an artifact due to spectral leakage from the zero-frequency bin, because 1) we used linear detrending to compute power spectra, and 2) difference-mode involves subtracting two conditions and the amount of spectral leakage should not depend strongly upon condition. The topography has a primary peak at electrode PO7, and a secondary peak at electrode AF3, perhaps extending to AF4. The temporal factor (thick solid line) is plotted with the group-averaged, difference-mode PSD at 1 Hz, for electrodes PO7 (thin solid line) and AF3 (thin dashed line). Overall, the temporal behavior of electrode PO7 and electrode AF3 are quite similar, although PO7 is larger for $t<0.5$ s. The difference in the temporal factors is that DM (1,1,1) reflects the late behavior that peaks around $t\approx 1.3$ s, while DM (1,1,2) reflects the early behavior that is confined to $t<0.5$ s. An explanation of how and why PCA separated these two temporal factors is given in the Discussion.

Discussion

Time–frequency analysis of multi-electrode EEG data in cognitive studies yields high-dimensional numerical results spanning space, time, frequency, conditions and subjects. There is a clear need to reduce and summarize these data, with the goal of isolating distinct neural processes involved in the task. Our initial attempt to extend the established technique of PCA-ANOVA to the frequency domain revealed three main shortcomings: 1) isolation of non-task-related differences between conditions (e.g., 60 Hz noise), 2) sensitivity to the precise definition of baseline power, and 3) sensitivity to the deletion of a single subject. We developed a new approach, called STAT-PCA, which advocates within-subject statistical testing followed by PCA. STAT-PCA was demonstrated to remedy all three of the short-comings of PCA-ANOVA, and yield components that have visible agreement with the group-averaged data. Upon close consideration, it makes intuitive sense that isolating task-related differences in single subjects improves the performance of PCA, because PCA always arranges the largest contributors to variance in its first few components. Only if the interesting data features are also the largest data features will PCA arrange them properly. Furthermore, STAT-PCA requires that a power difference reach significance in the single subject before it can contribute to the covariance matrix, while PCA-ANOVA considers all contributions that may or may not be significant within single subjects. For this reason, we propose that STAT-PCA is conceptually more rigorous than PCA-ANOVA. Finally, STAT-PCA permits the study of activity that is not only different between conditions, but also common to both conditions.

Sequential PCA produces factors that branch like a tree. The first PCA, in this case spectral, gives several factors. For each spectral factor, the second PCA, in this case spatial, gives one or more factors. For each spatial factor, the third PCA, in this case temporal, gives one or more factors. In this way, sequential PCA can tease apart features of the data that differ only at lower levels. An example of this is DM (1,1,1) and DM (1,1,2), which share spectral and spatial factors, but have different temporal factors. At first this separation may seem arbitrary, but it occurred presumably because of structure in the covariance matrix. The fact that this separation is visible in the group-averaged data provides corroborating evidence that this separation is valid. It is therefore a success of STAT-PCA to disambiguate these factors, and we interpret them as distinct physiological processes that likely have distinct interpretations in terms of cognitive processing as well.

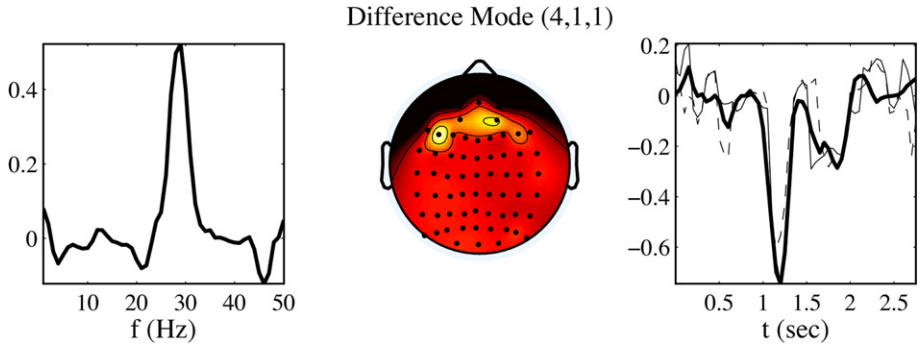

Fig. 6. STAT-PCA triplet DM (4,1,1). The temporal factor (thick solid line) is plotted along with the group-averaged, difference-mode PSD at 29 Hz, for electrodes F5 (thin solid line) and AF4 (thin dashed line).

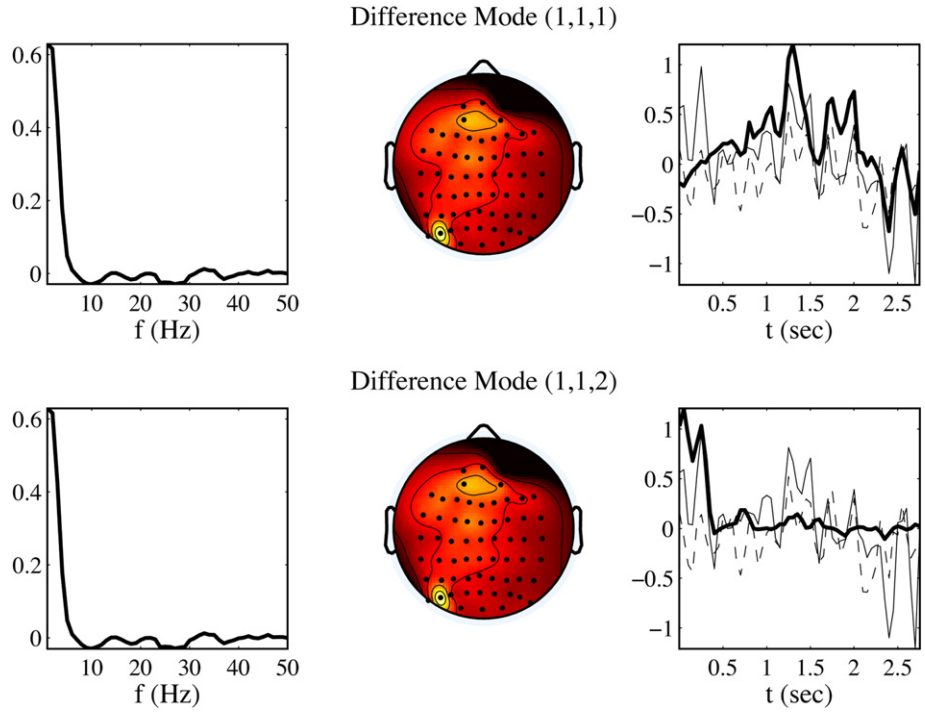

Fig. 7. STAT-PCA triplets DM (1,1,1) and DM (1,1,2). The temporal factor (thick solid line) is plotted with the group-averaged, difference-mode PSD at 1 Hz, for electrodes PO7 (thin solid line) and AF3 (thin dashed line).

Varimax rotation was applied to spectral, spatial, and temporal dimensions. By definition, it produces eigenvectors that are concentrated on a few elements, but those elements need not be proximal to each other. In the spectral domain, most factor loadings had a single, prominent peak, with one or more small sub-peaks. This situation is consistent with prior knowledge that task-related cortical oscillations tend to be relatively narrow-banded, and different bands tend not to overlap. In the spatial domain, Varimax rotation produces maps involving few electrodes, but these electrodes need not be adjacent, as seen in Figs. 6 and 7. In the temporal domain, the loadings tended to be less compact, reflecting sustained neural oscillations during the task. Despite this tendency, temporal factors may be compact, as seen in Fig. 7.

Unlike the standard method PCA-ANOVA, which requires multiple subjects as samples for the ANOVA, a major strength of the new method, STAT-PCA, is the ability to analyze single subjects. Because the goal of the present paper was to compare STAT-PCA with PCA-ANOVA, however, we applied the two methods on equal footing, focusing on group analysis. It might be supposed that, because PCA is a variance-based technique, multi-subject PCA would be more sensitive to inter-subject differences than commonalities. We argue against this possibility on several grounds. First, we showed stability of the factors under deletion of single subjects, thus no single subject (remembering that some subjects can be considered outliers) overly influences the results. Second, in Figs. 5–7 the high correlation between temporal factors (thick lines) and group-averaged power (thin lines) confirms that the STAT-PCA factors reflect the group-averaged behavior. Third, we have begun a follow up study in which we have done single-subject analysis using STAT-PCA, and have found preliminarily that the major factors that emerge for the group are visible in most of the single subjects. Beyond these points, group analysis provides advantages over single-subject analysis, because only in group analysis is the last (temporal) dimension submitted to PCA. As noted above, Fig. 7 shows how, for a particular spectral and spatial factor, the temporal PCA identified two temporal factors. In order to do this last (temporal) PCA, we had to use subjects as samples. When doing single-subject STAT-PCA, the last (temporal) factors must be obtained as the scores of the previous (spatial) PCA, thus there can be only one temporal factor for each spatial factor. For this reason, it would not have been possible to separate the two temporal factors shown in Fig. 7 in single subjects. Future work will investigate thoroughly the relationship between group and single-subject analyses, especially because the latter is necessary for clinical diagnosis.

A related approach, multi-way or parallel factor (PARAFAC) analysis, has also been applied to reduce space–time–frequency data (Miwakeichi et al., 2004). Because it operates on all three dimensions simultaneously, it avoids the issue of ordering in sequential PCA. Although PARAFAC is receiving much recent attention in the literature, its structure is such that each spectral factor is associated with only one spatial and temporal factor. It seems unlikely that PARAFAC would have separated the two processes shown in Fig. 7. Furthermore, it is often noted that PARAFAC solutions are unique, avoiding the rotation ambiguity in PCA. Practically speaking, however, the use of PARAFAC depends upon several choices, including the number of factors to retain, and constraints between factors in each dimension: orthogonality, positive-definiteness, and compactness. In this sense, the results of PARAFAC analysis are not unique, and more work is required to understand the effects of these constraints and the relationship of PARAFAC with PCA.

Beyond the internal consistencies of STAT-PCA that are the emphasis of this paper, at least one of the triplets found here is consistent with published findings in this task. DM (4,1,1) showed that electrodes F5 and AF3 oscillate in the range 20–35 Hz, and have nearly identical time course. First, this frequency band was found in this same task using intra-cranial electrodes (Slotnick et al., 2002). Second, electrode F5 was found previously by applying PCA-ANOVA to ERP analysis of the same data set (Brier et al., 2008). Third, the time duration (0.75–1.5 s) is the same as that of the ERP. Finding what appears to be the same neural process with ERP and time–frequency analysis gives further credibility to our approach developed here, and provides new ideas for how to clarify its functional role in semantic processing. Because electrodes F5 and AF3 oscillate at the

same frequency and share the same time course, we hypothesize that coherence analysis or phase synchrony analysis (Lachaux et al., 1999) will show these areas to be coupled in the frequency band 20–35 Hz.

In summary, STAT-PCA provides a basis for the reduction of the results of time–frequency analysis of multi-electrode EEG data into concise components that facilitate cognitive interpretation. It represents a paradigm shift for the integration of PCA with statistical testing, by advocating statistical tests in single subjects prior to PCA. In this way, PCA is relegated to a purely descriptive role. As a result of the statistical test occurring first, the factors retained do not need to be subjected to further statistical testing, which previously has been highly subjective. We conclude that STAT-PCA represents an extensible platform for the analysis of event-related spectral changes in cognitive experiments, as well as an adaptive platform for future developments that should include higher-dimensional (i.e., more than two-condition) experimental designs.

Acknowledgments

The authors thank Cliff Calley and Mary Katherine Reagor for assistance with data acquisition, Priya Xavier and Audrey Chang for assistance with pre-processing, Jeff Spence, Pat Carmack, and Wayne Woodward for discussions about spectral estimation, Gail Tillman, Mandy Maguire, and Joseph Dien for discussions about PCA, and Michael Motes for discussions about stability analysis,. This work was supported in part by the Berman Research Initiative, and the Departments of Radiology (Dr. Richard Briggs) and Internal Medicine (Dr. Robert Haley) at the University of Texas Southwestern Medical Center.

Appendix A

Implementation of PCA

In each step of the PCA, the time-varying PSD values are arranged in a matrix with variables denoting the PCA dimension, e.g., spectral, spatial, or temporal, and rows denoting the concatenation of the remaining variables, conditions (PCA-ANOVA only), and subjects. Given a data matrix X, where rows denote samples and columns denote variables, the first step of PCA is to subtract the column-means, i.e., for each column separately the mean across rows is subtracted:

$$Y \equiv X - \overline{X}.$$

(Because rows denote samples, the column-mean may be interpreted as the sample-mean for each variable.)

This mean-subtracted data matrix Y has a singular value decomposition:

$$Y - USV^T$$

where U is an orthogonal matrix with columns equal to the left eigenvectors, S is a diagonal matrix of singular values, and V is an orthogonal matrix with columns equal to the right eigenvectors. By definition, an orthogonal matrix satisfies $U^TU = 1$; the orthogonality of the matrix U is accomplished by the orthonormality of each column of U. The number of non-zero singular values is equal to the lesser of the number of samples or variables.

The covariance matrix is defined:

$$C \equiv Y^TY = \left[USV^T\right]^T\left[USV^T\right] = VS^2V^T$$

where the normalization factor (equal to the number of samples minus one) has been ignored in its definition, because an overall constant does not affect the decomposition. The last equality arises from the orthogonality of the matrix U. In the language of PCA, the right eigenvectors are called the factor loadings, and the elements of S^2 are called the eigenvalues.

The factor scores W are obtained by projecting the data onto the orthonormal basis comprised of the factor loadings:

$$W \equiv YV = US$$

where the second equality results from the SVD of Y. The factor scores are not merely the left eigenvectors, but include the singular values. In this way, the weight, i.e., the singular value, with which each eigenvector in U enters into the mean-subtracted data is included in the corresponding score vector.

Varimax rotation is applied to the factor loadings V, resulting in the rotated factor loadings:

$$V' = RV.$$

Varimax rotation preserves inner products between the eigenvectors in V, i.e., orthogonality and normalization, so R is an orthogonal matrix. Many applications of PCA consider only the factor loadings, so the effect of rotation on the factor scores is not considered. In the results presented here, the factor scores were derived by projecting the mean-subtracted data onto the rotated factor loadings:

$$W' = YV'.$$

Only in this way can the original data be considered to be comprised of the resulting factor scores and factor loadings:

$$W'V'^T = YV'V'^T = Y.$$

Keeping track of the rotation matrix **R** explicitly:

$$\begin{aligned} W'V'^T &= [YRV][RV]^T \\ &= YRVV^TR^T \\ &= YRR^T \\ &= Y \end{aligned}$$

where the third equality arises from the orthogonality of V, and the last equality arises from the orthogonality of R. In each step of the PCA, therefore, a subset of the factor loadings were retained and rotated. For each rotated factor loading, the corresponding factor score was computed by projecting the data matrix onto that loading. Each of these 'rotated' scores was reshaped to form a new data matrix for the next step of PCA.

References

Assaf, M., Calhoun, V.D., Kuzu, C.H., Kraut, M.A., Rivkin, P.R., Hart, J., Pearlson, G.D., 2006. Neural correlates of the object-recall process in semantic memory. Psychiatry Research. Neuroimaging 147, 115–126.

Bernat, E.M., Williams, W.J., Gehring, W.J., 2005. Decomposing ERP time–frequency energy using PCA. Clin. Neurophysiol. 116, 1314–1334.

Bokil, H., Purpura, K., Schoffelen, J.M., Thomson, D., Mitra, P., 2007. Comparing power spectra and coherences for groups of unequal size. J. Neurosci. Methods 159, 337–345.

Brier, M.R., Maguire, M.J., Tillman, G.D., Hart, J., Kraut, M.A., 2008. Event-related potentials in semantic memory retrieval. Journal of the Int. Neuropsychol. Soc. 14 (5), 815–822.

Cattell, R.B., 1966. The scree test for the number of factors. Multivariate Behav. Res. 1, 245–276.

Curry, S.H., Cooper, R., McCallum, W.C., Pocock, P.V., Papakostopoulos, D., Skidmore, S., Newton, P., 1983. The principle components of auditory target detection. In: Gaillard, A.W.K., Ritter, W. (Eds.), Tutorials on ERSP Research: Endogenous Components. North-Holland Publishing Company, Amsterdam, pp. 79–117.

Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent components analysis. J. Neurosci. Methods 134, 9–21.

Dien, J., 1998. Addressing misallocation of variance in principal components analysis of event-related potentials. Brain Topogr. 11 (1), 43–55.

Dien, J., Frishkoff, G.A., 2005. Principal component analysis of ERP data. In: Handy, T.C. (Ed.), Event-Related Potentials: A Methods Handbook. MIT Press.

Dien, J., Spencer, K.M., Donchin, E., 2003. Localization of the event-related potential novelty response as defined by principal components analysis. Cogn. Brain Res. 17, 637–650.

Dien, J., Beal, D.J., Berg, P., 2005. Optimizing principal components analysis of event-related potentials: matrix type, factor loading weighting, extraction, and rotations. Clin. Neurophysiol. 116, 1808–1825.

Dien, J., Khoe, W., Mangun, G.R., 2007. Evaluation of PCA and ICA of simulated ERPs: Promax vs. Infomax rotations. Hum. Brain Mapp. 28, 742–763.
Ding, M., Bressler, S.L., Yang, W., Liang, H., 2000. Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. Biol. Cybern. 83, 35–45.
Donchin, E., 1966. A multivariate approach to the analysis of event-related potentials. IEEE Trans. Biomed. Eng. 13 (3), 131–139.
Donchin, E., Heffley, E., 1979. Multivariate analysis of event-related potential data: a tutorial review. In: Otto, D. (Ed.), Multidisciplinary Perspectives in Event-related Potential Research. EPA 600/9-77-043). US Government Printing Office, Washington, DC, pp. 555–572.
Durka, P.J., Zygierewicz, J., Klekowicz, H., Ginter, J., Blinowska, K.J., 2004. On the statistical significance of event-related EEG desynchronization and synchronization in the time-frequency plane. IEEE Trans. Biomed. Eng. 51 (7), 1167–1175.
Ferree, T.C., 2006. Spherical splines and average referencing in scalp EEG. Brain Topogr. 19 (1/2), 43–52.
Ferree, T.C., Luu, P., Russell, G.S., Tucker, D.M., 2001. Scalp electrode impedance, infection risk, and EEG data quality. Clin. Neurophys. 112, 536–544.
Glorfeld, L.W., 1995. An improvement ion Horn's parallel analysis methodology for selecting the correct number of factors to retain. Educ. Psychol. Meas. 55, 377–393.
Hayton, J.C., Allen, D.G., Scarpello, V., 2004. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. Organ. Res. Methods 7, 191–205.
Horn, J.L., 1965. A rationale and test for the number of factors in factor analysis. Psychometrika 32, 179–185.
Junghofer, M., Elbert, T., Tucker, D.M., Braun, C., 1999. The polar average reference effect: a bias in estimating the head surface integral in EEG recording. Clin. Neurophysiology 110 (6), 1149–1155.
Lachaux, J.P., Rodriguez, E., Marinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Hum. Brain Mapp. 8, 194–208.
Kalcher, J., Pfurtscheller, G., 1995. Discrimination between phase-locked and non-phase-locked event-related EEG activity. Electroenceph and Clin. Neurophysiol. 94, 381–384.
Kayser, J., Tenke, C.E., 2003. Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation. Clin. Neurophysiol. 114, 2307–2325.
Miwakeichi, F., Martinez-Montes, E., Valdes-Sosa, P.A., Nishiyama, N., Mizuhara, H., Yamaguchi, Y., 2004. Decomposing EEG data into space–time–frequency components using Parallel Factor Analysis. NeuroImage 22, 1035–1045.
Mocks, J., Verleger, R., 1991. Multivariate methods in biosignal analysis: application of principal component analysis to event-related potentials. In: Weitkunat, R. (Ed.), Digital Biosignal Processing. Elsevier, Amsterdam, pp. 399–458.
Nunez, P.L., 1981. Electric Fields of the Brain. Oxford University Press.
Perrin, F., Pernier, J., Bertrand, O., Echallier, J.F., 1989. Spherical splines for scalp potential and current density mapping. Electroencephalogr. Clin. Neurophysiol. 72, 184–187.
Pfurtscheller, G., Lopes da Silva, F.H., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin. Neurophysiol. 110, 1842–1857.
Slotnick, S.D., Moo, L.R., Kraut, M.A., Lesser, R.P., Hart, J., 2002. Interactions between thalamic and cortical rhythms during semantic memory recall in human. Proc. Natl. Acad. Sci. 99 (9), 6440–6443.
Spencer, K.M., Dien, J., Donchin, E., 1999. A componential analysis of the ERP elicited by novel events using a dense electrode array. Psychophysiology 36, 409–414.
Spencer, K.M., Dien, J., Donchin, E., 2001. Spatiotemporal analysis of the late ERP responses to deviant stimuli. Psychophysiology 38, 343–358.
Tenke, C.E., Kayser, J., 2005. Reference-free quantification of EEG spectra: combining current source density (CSD) and frequency principal components analysis (fPCA). Clin. Neurophysiol. 116, 2826–2846.
Thomson, D.J., Chave, A.D., 1991. Jackknifed error estimates for spectra, coherences, and transfer functions. In: Haykin, S. (Ed.), Advances in Spectrum Estimation. Prentice Hall, pp. 58–113.
Truccolo, W.A., Ding, M., Knuth, K.H., Nakamura, R., Bressler, S.L., 2002. Trial-to-trial variability of cortical evoked responses: implications for the analysis of functional connectivity. Clin. Neurophysiol. 113 (2002), 206–226.
Turner, N.E., 1998. The effect of common variance and structure pattern on random data eigenvalues: implications for the accuracy of parallel analysis. Educ. Psychol. Meas. 58, 541–568.
Welch, P.D., 1967. The use of fast Fourier transform for estimation of power spectra: a method based on time-averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. 15 (2), 70–73.
Wood, C.C., McCarthy, G., 1984. Principal component analysis of event-related potentials: simulation studies demonstrate misallocation of variance across components. Electroenceph and Clin. Neurophysiol. 59, 249–260.
Zhu, M., Ghodsi, A., 2006. Automatic dimensionality selection from the screen plot via the use of profile likelihood. Comput. Stat. Data Anal. 51, 918–930.
Zwick, W.R., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. Psychol. Bull. 99 (3), 432–442.